An Xrootd Italian Federation

2014 J. Phys.: Conf. Ser. 513 042013

(http://iopscience.iop.org/1742-6596/513/4/042013)

Journal of Physics: Conference Series **513** (2014) 042013 doi:10.1088/1742-6596/513/4/042013

An Xrootd Italian Federation.

T.Boccali[3], G.Donvito[2], D.Diacono[2], G.Marzulli[2], A.Pompili[4], G.Della Ricca[5], E.Mazzoni[3], S.Argiro[6], D.Gregori[7], C.Grandi[7], D.Bonacorsi[8], L.Lista[9], F.Fabozzi[10], L.M.Barone[11], A.Santocchia[12], H. Riahi[13], A.Tricomi[14], M.Sgaravatto[15], G.Maron[1]

[1] INFN Laboratori di Legnaro
[2] INFN Sezione di Bari
[3] INFN Sezione di Pisa
[4] Politecnico di Bari
[5] Universita di Trieste
[6] Universita di Torino
[7] INFN CNAF
[8] Universita di Bologna
[9] INFN Sezione di Napoli
[10] Universita di Napoli Federico II
[11] Universita di Roma La Sapienza
[12] Universita di Perugia
[13] INFN Sezione di Perugia
[14] Universita di Catania
[15] INFN Sezione di Padova

E-mail: giacinto.donvito@ba.infn.it

Abstract. The Italian community in CMS has built a geographically distributed network in which all the data stored in the Italian region are available to all the users for their everyday work. This activity involves at different level all the CMS centers: the Tier1 at CNAF, all the four Tier2s (Bari, Rome, Legnaro and Pisa), and few Tier3s (Trieste, Perugia, Torino, Catania, Napoli, ...). The federation uses the new network connections as provided by GARR, our NREN (National Research and Education Network), which provides a minimum of 10 Gbit/s to all the sites via the GARR-X[2] project. The federation is currently based on Xrootd[1] technology, and on a Redirector aimed to seamlessly connect all the sites, giving the logical view of a single entity. A special configuration has been put in place for the Tier1, CNAF, where ad-hoc Xrootd changes have been implemented in order to protect the tape system from excessive stress, by not allowing WAN connections to access tape only files, on a file-by-file basis. In order to improve the overall performance while reading files, both in terms of bandwidth and latency, a hierarchy of xrootd redirectors has been implemented. The solution implemented provides a dedicated Redirector where all the INFN sites are registered, without considering their status (T1, T2, or T3 sites). An interesting use case were able to cover via the federation are disk-less Tier3s. The caching solution allows to operate a local storage with minimal human intervention: transfers are automatically done on a single file basis, and the cache is maintained operational by automatic removal of old files.

1. Introduction

Italy participates to the CMS Collaboration with a large number of physicists (roughly 13% of the total), distributed in 16 Institutions, and coordinated by INFN (National Institute for

Nuclear Physics). The analysis activities are supported via a number of sites, participating to the CMS Distributed Computing Project. Data custodiality and central processing activities are mostly performed at the Tier1 at CNAF (Bologna), while analysis activities are carried out at the four Tier2 sites (Pisa, Bari, Rome, Legnaro/Padova). A number of smaller sites (Tier3s), mostly financed outside INFN budget, via local projects or grants, also exist and is used by local communities for analysis purposes; these sites are generally quite rich in CPUs (order of 100 or more), and sometimes offer sizeable storage resources. In the latter case, the main differences between Tier2s and Tier3s is the funding model, the obligations towards the experiment (none, for Tier3s), and the quality of the local infrastructure.

The use of remote data access (via the Xrootd protocol, currently in use by CMS), hase been enabled to serve different purposes:

(i) use our Tier1-Tier2s backbone to directly serve data to the Italian physicists, from their local resources;

(ii) have a complete set of all the physics enabling data and simulations in the Italian centers, where we can guarantee bandwidth to the analyzers;

(iii) allow data sitting in our minor sites (the Tier3s) to be distributed among collaborators, without making it global inside CMS.

The first use case enters into the general CMS Xrootd Federation effort, described elsewhere[3]. The second and third are outside the usual CMS use cases and will be discussed in details.

2. Towards an Italian Xrootd Federation

Recent improvements in the networking technology have allowed for a partial paradigm shift in the CMS analysis model. Two factors should be taken into account:

- The Italian NREN (GARR) is now able to provide INFN Tiers with a geographical connectivity in excess of 10 Gbit/s to most Italian institutions supporting CMS. In particular, out Tier1 is connected at 40 Gbit/s, Tier2s vary between 10 an 20 Gbit/s, and Tier3s at 10 Gbit/s (apart from few cases still at 2 Gbit/s);
- The storage deployed at Tier1 and Tier2s, available for analysis activities, exceeds 6 PB; this makes the placing of all interesting recent data and Monte Carlo feasible within Italian boundaries.

A model in which computational tasks are not limited by data locality has become feasible, with (Italian bound) geographical networking comparable to local WAN links, and certainly sufficient to saturate the CPUs at typical Tier3s. Indeed, the CMS Collaboration has started to allow for direct access to remote files in specific cases, the Tier3s being one of these. A typical analysis workflow can run at Tier3, accessing data from Tier1 and Tier2s, and saving results locally.

In CMS, Tier3s have access to virtually all the data via the Xrootd Federation, whose entry point in Europe is the regional Redirector in Bari. All the European Tier2s (and some Tier1s) subscribe to this Redirector, and allow for remote file serving. The structure of the Redirector, however, is that of a round robin connection: if the same file is available (for example), in Pisa and London, it will be served from both location with a 50% probability, irrespective of the source location and networking specificities.

By building an Italian Xrootd Federation, we wanted to make sure to avoid exiting from GARR guaranteed network boundaries as much as possible. The solution has been identified in setting up a specific Italian Xrootd Redirector (`xrootd-it.ba.infn.it`), which allows for a more controlled traffic routing. This Redirector links in fallback to the European Regional Redirector.

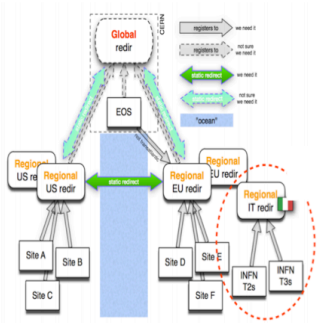

Figure 1. INFN in Italy has a dedicated Xrootd Redirector which all the Italian resources are registered to

In this configuration, depicted in Figure 1, the logic for serving a file is the following:

(i) a file is searched locally beforehand (this is outside the Xrootd chain);

(ii) if not available, the file is requested to the Italian Redirector; if positively found, it is accessed;

(iii) if not available on the Italian Redirector, the query is forwarded to the European Redirector, which searches for it in Europe; if positively found, it is accessed;

(iv) if not available even on the European Redirector, the query scales up to the Global Redirector, which has knowledge of all the files available vis Xrootd in CMS as a whole. If still not there, a file open error is issued.

In this way, given the large share of data available in Italian Tier1 and Tier2s, the need to search beyond national level is estimated at the percent level.

There is another advantage by having an isolated Italian-level Redirector as the first level in the hierarchy. Tier3s do have storage and CPUs resources, but usually of a lower quality with respect to that of bigger sites. Hence, having a standard Tier3 joining the CMS Xrootd Federation is risky, since there is no limit of the inbound traffic its storage would need to handle. Instead, we can quite safely serve Tier3 storage via Xrootd if we are sure we can maintain the traffic under better control, as it happens allowing for connections only from Italian sites. This on one hand decreases the load, on the other allows for fast actions in case a troubled situation is reached. We hence decided to have our Tier3s joining the Italian CMS Xrootd Federation, without joining the Global Federation. A collaborator in Italy can hence directly use remote data (or more probably ntuples) from the Tier3, simplifying common analysis efforts, on top of what the Tier1 and Tier2s already offer.

Preliminary tests have been run using the Italian federation. In all of them a site compares software (analysis on CMS data/MC and on precooked ntuples) speed by accessing data:

- locally;
- from an Italian Tier2;
- from an Italian Tier3.

Results for standard analysis tasks show a performance degrade (measured as loss in CPU efficiency) by 5-10%, perfectly acceptable given the added number of resources which can be accessed in this way. Analysis activity at a Tier3, with data accessed via the federation, has become the standard for Italian physicists in CMS.

3. Tier3 Xrootd specific settings

An interesting use case is that of "small" Tier3s, with very limited local storage resources. These sites could operate as disk-less CMS resources, by accessing data directly via the federation. A better optimization has been found with the installation of Xrootd caches at their boundaries. When used for end analysis tasks, indeed, these sites have been seen accessing over and over the same data, probably due to local analysis optimization tests. An Xrootd cache, even if of limited size (tens of TBs), allows all the optimization work to happen locally, after a first WAN transfer. To achieve this, a tape-like setup has been installed in these sites: when a file is not available on the local Xrootd, an attempt to recall it via the tape Xrootd daemon is made; tape calls are redirected to local transfers via `xrdcp`[4] from the federation. The files are thus copied locally, and accessed from local storage from that point onwards. Xrootd also allows for a transparent handling of the cache, with a `purge` daemon which makes sure the cache remains usable over time.

4. INFN-CNAF Tier1 Xrootd Setup

A major ingredient in the setup is the availability of Tier1 data via the federation. The CNAF Tier1, uses a storage solution based on a large GPFS[5] installation (exceeding 10 PB of disk), integrated with a tape backend using TSM[6]. The two systems are integrated via the in-house developed GEMSS[7] system.

Within the GPFS file system we can distinguish between two types of files:

- the class called T1D0, which are files on tape which need to be recalled in a buffer area in case of access;
- the class called T0D1, which are files resident on disk.

GPFS integrates all the files in a seamless way, by showing "stub files" for those files which are not resident on disk. Hence, GPFS always shows all the files, irrespective of their direct availability on disk. This creates a potential problem: a standard, POSIX based, Xrootd server would publish to the upstream Redirectors all the files present on the filesystem; remote clients requests would then trigger tape recalls in a not predictable way thus harming the tape drives, which are setup to handle a limited number of file recall per hour.

A solution which consists in hiding from the Xrootd Redirectors the files which are only present on the tape backend has been designed and implemented with Xrootd developers. Files not on disk (recognized via specific fields in the `stat` command) are not communicated to the Redirectors, via a patch in the `cmsd` daemon. In this way the files are unknown to the Global Redirector, and will not be served.

The only way to access those file is via a direct stage-in command issued before the Xrootd attempt is tried. Since this can be a limiting factor in the case when we want to be able to trigger automatic recall, a second mechanism has been put in place: direct requests to the CNAF Xrootd servers (not via Redirector) are allowed also for tape-only files, but are not triggering direct recalls; these are instead queued in the GEMSS system, which makes sure that recalls are done safely for the tape drives, and for example joining requests for files on the same tape.

In production there are currently four Xrootd - gridftp servers a schematic layout is shown in Figure 2.

Previous experiences with older Xrootd versions showed large increases in the CPU load of each server. For this reason, we have carried out preliminary tests to simulate conditions of high data traffic. With the availability of four servers gridftp in production, each with 10 Gb/s Ethernet network interface, has activated the Xrootd on one of these and have been read 30TB of data via Xrootd.

To ensure the proper functioning of the service we have also developed some Nagios controls that make copies (functionality test) using the command `xrdcp`[4] from the CMS storage area to

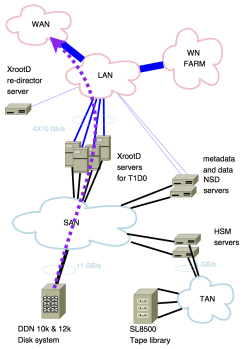

Figure 2. CMS layout, each server is a Xrootd - gridftp server in production, with 10 Gb/s ethernet connection and 8 Gb/s Fibre Channel Port to access the Storage Area Network. The data that is stored only on tape are not published by the process CMSD, that is one of the two daemon of XRootD server.

the local disk of the dedicated server where all the required software is installed. These checks are carried out at intervals of 4 hours. More frequent checks shall verify the status of services and in case of error Nagios sends email notification.

5. DNS High Availability

The availability of Xrootd Redirectors is critical in the system, since they are at the first services users have to contact to perform analysis activities. We are implementing an high availability solution for the Xrootd Redirector, based on the DNS response. The solution will check the availability of Xrootd service and reconfigure the DNS accordingly: in this way if the main Xrootd Redirector goes offline, the DNS will be reconfigured with a different machine IP that has the services configured and in stand-by. In this way the Xrootd client will always obtain a valid IP for an host running Xrootd service.

6. Conclusions

The Italian Xrootd Federation is currently in place, with the majority of Tier3s already attached. The solution is already producing good results in terms of data availability and site reliability. The possibility to exploit the Xrootd infrastructures also at the Tier3 level is increasing the possibility for the end users to share data among colleagues that are working within the same physics groups. The availability of also the file hosted at the Tier1 on the other hand greatly increase the capabilities of processing any kind of data tiers.

The work has been partially funded under contract 20108T4XTM of 'Programmi di Ricerca Scientifica di Rilevante Interesse Nazionale (Italy)'.

Bibliography

[1] A. Hanushevsky at Al., *Scalla: Structured Cluster Architecture for Low Latency Access*, published in Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2012 IEEE 26th International.
[2] http://www.garr.it/b/eng
[3] K. Bloom et Al.,*CMS Use Of a Data Federation*, this conference.
[4] Manual Available Online `http://linux.die.net/man/1/xrdcp`
[5] IBM, *GPFS: A Shared-Disk File System for Large Computing Clusters*, Published in Proceeding FAST '02 Proceedings of the 1st USENIX Conference on File and Storage Technologies.
[6] IBM, *http://public.dhe.ibm.com/software/in/tivoli/ESG_White_Paper_IBM_TSM_6_Apr_09.pdf.*
[7] D. Bonaccorsi et Al., *The Grid Enabled Mass Storage System (GEMSS): the Storage and Data management system used at the INFN Tier1 at CNAF*, 2012 J. Phys.: Conf. Ser. 396 042051.